# Introduction

## Overview

* Analysis of research papers can give a lot of insights about software resources and their dependency.
* In a scientific research different kinds of input resources are used. One of such input is a software.
* Used resources in a research are typically mentioned in a citation. Citation practices of formal articles in a research are matured and various citation styles exist. Even if principles for formal citation of a software has already been put out, most scientists are not properly citing resources.
* Surprisingly, sometimes researchers do not mention the type of software they used entirely or mention it with a rather vague abbreviation and just talk about the results they have obtained.
* As long as software is mentioned using formal methods, like RRID, it is possible to perform citation analysis using regular expressions which can be constructed to capture the pattern of citation.
* Though regular expression based analysis can give basic insights about the software citation it has limitations because:
  + Not so many authors use formal citation of software, like RRIDs
  + Even if scientists use formal citations, they may fail to properly follow the guidelines. For example, some authors tend to ignore the RRID-part and that creates an ambiguity by it self that it is not possible to know weather the author is actually making a software citation or it is completely something else.
  + Rule based method fails to capture context information and ignores dependencies. It is not possible to be sure about the authors intention whether or not using a software citation.
* At the same time pattern based analysis, like using regX, is not suitable to extract information about software citation, for instance the particular use of a software, especially when a software mention statement lacks any form of formality where the information is concealed in a natural language description.
* Therefore it is required to automatically extract the purpose of software use in scientific literatures. This might help to answer questions like:
  + What type of software is being frequently used for what purpose in a specific area of research? This also allows to find an answer further question like what is the most common technique researchers follow when trying to solve a given research problem in a given domain )
* Previous attempts to automatically extract information using machine learning techniques, specifically supervised machine learning technique, about the software use purpose was constrained mainly because of lack of ground truth data. But this time, with the advent of SoMeSci, it is possible to do so.

## Problem statement

## Objectives of the research

This work has the following objectives:

* List down the purpose of software usage in a research in a hierarchical manner.
* To extend SoMeSci with a manual annotation of purpose of software usage.
* To find a classifier and optimize results.

# Literature review

## The role of Software in Scientific research

### Introduction

* The use of software is very popular in a modern research [3 , 4, 14].
* More and more scientists are developing a software as part of their scientific work according to a survey done in 2008 [1, 2, 6] :
  + nearly 45% of scientists spend more time developing a software as part of their research work than five years ago.
  + 38% of researchers spend at least 20% of their time developing a software.
  + over 90% of scientists agree that software is important for their research and
  + nearly 70% claim that their research directly depends on a use of a software.
* Popularity of a given software can be domain specific i.e. in each domain there can be a popular software. Similarly, some areas of research like economics does not heavily rely on software compared to other fields like physics.
* There is a clear distinction between a scientific software and regular software [4]. A *scientific software* often is developed by researchers who have domain specific knowledge [3] as part of their research and such a software is intended to help understand the research problem [2]. In contrast, a *regular software* is refers to any software that does not heavily rely on domain specific knowledge for its development and can be done by any self-taught software enthusiast. *In this paper the term software refers to a scientific software* (is this scope valid?).

### General roles of software in a research

* Role of software #1: Software determines the quality and reliability of the research outcome [2].

A use of quality software produces better science(Goble, 2014)

Example: retrospective discovery of an error , logic error, in a code of a software renders the entire outcome of a research useless [5]. Following such events several retractions of research papers have been observed in the past [3, 6, 8].

* Role of software #2: Software can be used to validate results of a given research work.

Quality software produces a consistent and repeatable research output [9]. It also makes it easy for validation of the research results by other experimenters using the same software. A list of best practices [3, 9] from software engineering assist the development of quality software.

* Even though a software plays critical role in a modern research, most of the time research software developers will not get enough credits. Scientific software is also usually not considered as a research output (Hafer & Kirkpatrick, 2009).

### Concrete roles of software in a research

More concrete but general examples for roles of a software in a research are modelling , simulation and data analysis [6].

* Concrete role - Data Analysis

For instance, the Large Hadron Collider (LHC) facility at CERN uses a software consisting of over five million lines of code for collection and analysis of terabytes of data [4].

* Concrete role – Prediction from Data

For example, historical weather data can be used to make predictions about future temperature variations [4].

* Concrete role – Simulation

In a research scientists may use simulation software to model and mimic the behavior of real world phenomena where observation in a physical world is costly , disastrous or impossible[7].

## Literature review on classification of software usage purpose

* Sometimes scientists use several software together in their research[12].

This because that each software has one or more purpose that serves to explore the research problem and help to find an answer. Example: a scientist might use a data analysis software together with modelling software.

* A software can be used for various purposes during a research.

For example: to perform Data analysis (might refer to inspecting, cleaning, transformation, and modelling a data). The main purpose of data analysis is to extract meaningful information from a data that will help to make businesses operate more effectively [13].